Biometrika (2022), **109**, 1, *pp.* 123–136
Printed in Great Britain
doi: 10.1093/biomet/asab005
Advance Access publication 8 September 2021

Large-scale model selection in misspecified generalized linear models

By EMRE DEMIRKAYA

Department of Business Analytics and Statistics, University of Tennessee, 453 Haslam Business Building, Knoxville, Tennessee 37996, U.S.A.
edemirka@utk.edu

YANG FENG

Department of Biostatistics, School of Global Public Health, New York University, 715 Broadway, New York, New York 10003, U.S.A.
yang.feng@nyu.edu

PALLAVI BASU

Indian School of Business, Gachibowli, Hyderabad Telangana 500 111, India
plvibasu.work@gmail.com

and JINCHI LV

Data Sciences and Operations Department, University of Southern California, 3670 Trousdale Parkway, Los Angeles, California 90089, U.S.A.
jinchilv@marshall.usc.edu

Summary

Model selection is crucial both to high-dimensional learning and to inference for contemporary big data applications in pinpointing the best set of covariates among a sequence of candidate interpretable models. Most existing work implicitly assumes that the models are correctly specified or have fixed dimensionality, yet both model misspecification and high dimensionality are prevalent in practice. In this paper, we exploit the framework of model selection principles under the misspecified generalized linear models presented in Lv & Liu (2014), and investigate the asymptotic expansion of the posterior model probability in the setting of high-dimensional misspecified models. With a natural choice of prior probabilities that encourages interpretability and incorporates the Kullback–Leibler divergence, we suggest using the high-dimensional generalized Bayesian information criterion with prior probability for large-scale model selection with misspecification. Our new information criterion characterizes the impacts of both model misspecification and high dimensionality on model selection. We further establish the consistency of covariance contrast matrix estimation and the model selection consistency of the new information criterion in ultrahigh dimensions under some mild regularity conditions. Our numerical studies demonstrate that the proposed method enjoys improved model selection consistency over its main competitors.

Some key words: Bayesian principle; Big data; High dimensionality; Kullback–Leibler divergence; Model misspecification; Model selection; Robustness.

1. INTRODUCTION

With the rapid advances in modern technology, big data of unprecedented size, such as genetic and proteomic data, fMRI and functional data, and panel data in economics and finance, are frequently encountered in many contemporary applications. In these applications, the dimensionality p can be comparable to or even much larger than the sample size n. A key assumption that often makes large-scale learning and inference feasible is the sparsity of signals, meaning that only a small fraction of covariates contribute to the response when p is large relative to n. High-dimensional modelling with dimensionality reduction and feature selection plays an important role in these problems (e.g., Fan & Lv, 2010; Bühlmann & van de Geer, 2011; Fan & Lv, 2018). A sparse modelling procedure typically produces a sequence of interpretable candidate models, each involving a possibly different subset of covariates. An important question is how to compare different models in high dimensions when the models are possibly misspecified.

The problem of model selection has been studied extensively by many researchers over the past several decades. Well-known model selection criteria include the Akaike information criterion, AIC (Akaike, 1973, 1974), and the Bayesian information criterion, BIC (Schwarz, 1978), where the former is based on the Kullback–Leibler divergence principle of model selection and the latter on the Bayesian principle of model selection. A great deal of work has been devoted to understanding and extending these model selection criteria to different model settings (see, e.g., Bozdogan, 1987; Foster & George, 1994; Konishi & Kitagawa, 1996; Ing, 2007; Chen & Chen, 2008; Chen & Chan, 2011; Liu & Yang, 2011; Ninomiya & Kawano, 2016; Eguchi, 2017; Hsu et al., 2019). Fong & Holmes (2020) studied the links between cross-validation and Bayesian model selection. The connections between the AIC and cross-validation have been investigated by Stone (1977), Hall (1990) and Peng et al. (2013) in various contexts. In particular, Fan & Tang (2013) showed that classical information criteria, such as the AIC and BIC, can no longer be consistent for model selection in ultrahigh dimensions and proposed the generalized information criterion, GIC, for tuning parameter selection in high-dimensional penalized likelihood, when models are correctly specified. See also Barber & Candès (2015), Bühlmann & van de Geer (2015), Candès et al. (2018), Shah & Bühlmann (2018) and Fan et al. (2019, 2020) for some recent work on high-dimensional inference for feature selection.

Most existing work on model selection and feature selection makes an implicit assumption that the model under study is correctly specified or of fixed dimensions. Given the practical importance of model misspecification, White (1982) laid out a general theory of maximum likelihood estimation in misspecified models for the case of fixed dimensionality and independent and identically distributed observations. Cule et al. (2010) also studied the maximum likelihood estimation of a multi-dimensional log-concave density in the case where the model is misspecified. Recently, Lv & Liu (2014) investigated the problem of model selection with model misspecification and derived asymptotic expansions of both the Kullback–Leibler divergence and Bayesian principles in misspecified generalized linear models, leading to the generalized Akaike information criterion, GAIC, and generalized Bayesian information criterion, GBIC, for the case of fixed dimensionality. A specific form of prior probabilities motivated by the Kullback–Leibler divergence principle led to the generalized Bayesian information criterion with prior probability, GBIC_p. As contemporary big data applications often feature both model misspecification and high dimensionality, an important question is how to characterize their impact on model selection. The present paper aims to provide some partial answers to this question.

Let us first gain some insights into the challenges of the aforementioned problem by considering a motivating example. Assume that the response Y depends on the covariate vector $(X_1, \ldots, X_p)^{\mathrm{T}}$ through the functional form $Y = f(X_1) + f(X_2 - X_3) + f(X_4 - X_5) + \varepsilon$, where $f(x) = x^3/(x^2 + 1)$ and the other settings are as described in § 4.2. Let the sample size be $n = 200$ and vary the dimensionality p from 100 to 3200. Without any prior knowledge of the true model structure, we take the linear regression model

$$y = Z\beta + \varepsilon \tag{1}$$

as the working model and apply some information criteria to hopefully recover the oracle working model; here y is an n-dimensional response vector, Z is an $n \times p$ design matrix, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-dimensional regression coefficient vector, and ε is an n-dimensional error vector. Following Candès et al. (2018), we define the oracle working model $\mathfrak{M}_0$ to be the Markov blanket for Y, i.e., $\mathfrak{M}_0$ is the smallest subset of indices such that Y is independent of $X_{\mathfrak{M}_0^{\mathrm{c}}}$ conditional on $X_{\mathfrak{M}_0}$; see Lauritzen (1996) and Pearl (2014). In this example, the oracle working model consists of the first five covariates. When $p = 100$, the traditional AIC and BIC, which ignore model misspecification, tend to select a model of size greater than five. In contrast, $\mathrm{GBIC}_{\mathrm{p}}$ of Lv & Liu (2014) selects the oracle working model around 60% of the time. However, when p is increased to 3200, these methods fail to select such a model with significant probability and the prediction performance of the selected models deteriorates. This motivates us to study the problem of model selection in high-dimensional misspecified models. Our new method, in contrast, can recover the oracle working model with significant probability in this challenging scenario.

The main contributions of this paper are threefold. First, we derive the asymptotic expansion of the posterior model probability in high-dimensional misspecified generalized linear models, which involves delicate and challenging technical analysis. Motivated by the asymptotic expansion and a natural choice of prior probabilities that encourages interpretability and incorporates Kullback–Leibler divergence, we propose a method called the high-dimensional generalized Bayesian information criterion with prior probability, for large-scale model selection with misspecification. Second, our work provides rigorous theoretical justification of the covariance contrast matrix estimator that incorporates the effect of model misspecification and is crucial for practical implementation. Such an estimator is shown to be consistent in the general setting of high-dimensional misspecified models. Third, we establish the model selection consistency of our new information criterion in ultrahigh dimensions under some mild regularity conditions. In particular, our work provides important extensions of the studies in Lv & Liu (2014) and Fan & Tang (2013) to the cases of high dimensionality and model misspecification, respectively. The aforementioned contributions make our work distinct from other studies on model misspecification, such as those of Bühlmann & van de Geer (2015), Hsu et al. (2019) and Shah & Bühlmann (2018). Since Lv & Liu (2014) is closely related to the present paper, we reiterate the main differences between these two works. First, the study in Lv & Liu (2014) focused on fixed dimensionality, so our model selection criterion differs in how it penalizes the model complexity, as discussed in § 2.2. Although both criteria rely on estimation of the covariance contrast matrix, the consistency result of the covariance contrast matrix estimator in Lv & Liu (2014) does not allow model misspecification, whereas we establish in § 3.3 the consistency of the estimator for the covariance contrast matrix even under model misspecification. Finally, in light of the new consistency result, we further provide a model selection consistency theorem for our model selection criterion, a result that was missing from Lv & Liu (2014).

2. LARGE-SCALE MODEL SELECTION WITH MISSPECIFICATION

2.1. *Model misspecification*

The main focus of this paper is the investigation of ultrahigh-dimensional model selection with model misspecification in which the dimensionality p can grow nonpolynomially with the sample size n. Let Z be the $n \times p$ design matrix with all available covariates. We let $\mathfrak{M}$ denote an arbitrary subset of size d of the p available covariates and let $X = (x_1, \ldots, x_n)^{\mathrm{T}}$ denote the corresponding $n \times d$ fixed design matrix given by the covariates in model $\mathfrak{M}$. Assume that conditional on the covariates in model $\mathfrak{M}$, the response vector $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ has independent components and each Y_i follows distribution $G_{n,i}$ with density $g_{n,i}$, where all the distributions $G_{n,i}$ are unknown to us in practice. Denote by $g_n = \prod_{i=1}^n g_{n,i}$ the product density and G_n the corresponding true distribution of the response vector Y.

Since the collection of true distributions $\{G_{n,i}\}_{1 \leqslant i \leqslant n}$ is unknown to practitioners, one often chooses a family of working models to fit the data. A popular class of working models is the family of generalized linear models (McCullagh & Nelder, 1989) with a canonical link and natural parameter vector $\theta = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$ with $\theta_i = x_i^{\mathrm{T}}\beta$, where x_i is a d-dimensional covariate vector and $\beta = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$ is a d-dimensional regression coefficient vector. Let $\tau > 0$ be the dispersion parameter. Then under this working model, the conditional density of response y_i given the covariates in model $\mathfrak{M}$ is assumed to take the form

$$f_{n,i}(y_i) = \exp\{y_i\theta_i - b(\theta_i) + c(y_i, \tau)\}, \tag{2}$$

where $b(\cdot)$ and $c(\cdot\,, \cdot)$ are some known functions, with $b(\cdot)$ being twice continuously differentiable and $b''(\cdot)$ bounded away from 0 and ∞. Here F_n denotes the corresponding distribution of the n-dimensional response vector $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ with the product density $f_n = \prod_{i=1}^n f_{n,i}$, assuming the independence of components. Since the generalized linear model is chosen by the user, the working distribution F_n can generally be different from the true unknown distribution G_n.

For the generalized linear model in (2) with natural parameter vector θ, let us define two vector-valued functions $b(\theta) = \{b(\theta_1), \ldots, b(\theta_n)\}^{\mathrm{T}}$ and $\mu(\theta) = \{b'(\theta_1), \ldots, b'(\theta_n)\}^{\mathrm{T}}$, as well as a matrix-valued function $\Sigma(\theta) = \mathrm{diag}\{b''(\theta_1), \ldots, b''(\theta_n)\}$. Basic properties of generalized linear models give $E(y) = \mu(\theta)$ and $\mathrm{cov}(y) = \Sigma(\theta)$ with $\theta = X\beta$. The product density of the response vector y can be written as

$$f_n(y; \beta, \tau) = \prod_{i=1}^n f_{n,i}(y_i) = \exp\left\{y^{\mathrm{T}}X\beta - 1^{\mathrm{T}}b(X\beta) + \sum_{i=1}^n c(y_i, \tau)\right\}, \tag{3}$$

where 1 represents the n-dimensional vector with all components equal to the scalar 1. Since the family of generalized linear models is only our working model, (3) results in the quasi-loglikelihood function (White, 1982)

$$\ell_n(y; \beta, \tau) = \log f_n(y; \beta, \tau) = y^{\mathrm{T}}X\beta - 1^{\mathrm{T}}b(X\beta) + \sum_{i=1}^n c(y_i, \tau). \tag{4}$$

Hereafter we treat the dispersion parameter τ as a known parameter and focus on our main parameter of interest, β. Whenever confusion is unlikely, we will slightly abuse notation and drop the functional dependence on τ.

The quasi maximum likelihood estimator for the parameter vector β in our working model (2) is defined as $\hat{\beta}_n = \arg\max_{\beta \in \mathbb{R}^d} \ell_n(y, \beta)$, which is the solution to the score equation

$$\Psi_n(\beta) = \partial\ell_n(y, \beta)/\partial\beta = X^{\mathrm{T}}\{y - \mu(X\beta)\} = 0. \tag{5}$$

For the linear regression model with $\mu(X\beta) = X\beta$, this score equation becomes the familiar normal equation $X^{\mathrm{T}}y = X^{\mathrm{T}}X\beta$. Such a vector β is called the quasi maximum likelihood estimator when the model is misspecified. Hereafter, for simplicity we call β the maximum likelihood estimator, since we do not know whether or not the model is misspecified in practice. The Kullback–Leibler divergence (Kullback & Leibler, 1951) between our working model F_n and the true model G_n is defined as $I\{g_n;f_n(\cdot\,,\beta)\} = E\{\log g_n(Y)\} - E\{\ell_n(Y,\beta)\}$, with the response vector Y following the true distribution G_n. As in Lv & Liu (2014), we consider the best working model in the sense that it is closest to the true model under the Kullback–Leibler divergence. Such a model has parameter vector $\beta_{n,0} = \arg\min_{\beta\in\mathbb{R}^d} I\{g_n;f_n(\cdot\,,\beta)\}$, which solves the equation

$$X^{\mathrm{T}}\{E(Y) - \mu(X\beta)\} = 0. \tag{6}$$

We see that (6) is simply the population version of the score equation (5).

Following Lv & Liu (2014), we introduce two matrices, the Fisher information in outer product form and in Hessian form. These matrices play a key role in model selection with model misspecification. Under the true distribution G_n, we have $\mathrm{cov}(X^{\mathrm{T}}Y) = X^{\mathrm{T}}\mathrm{cov}(Y)X$. Computing the score equation at $\beta_{n,0}$, the Fisher information matrix in outer product form is

$$B_n = \mathrm{cov}\{\Psi_n(\beta_{n,0})\} = \mathrm{cov}(X^{\mathrm{T}}Y) = X^{\mathrm{T}}\mathrm{cov}(Y)X \tag{7}$$

with $\mathrm{cov}(Y) = \mathrm{diag}\{\mathrm{var}(Y_1),\ldots,\mathrm{var}(Y_n)\}$ by the independence assumption and under the true model. Under the working model F_n, we have $\mathrm{cov}(X^{\mathrm{T}}Y) = X^{\mathrm{T}}\Sigma(X\beta)X$. The Fisher information matrix in Hessian form is defined by

$$A_n(\beta) = \frac{\partial^2 I\{g_n;f_n(\cdot\,,\beta)\}}{\partial\beta^2} = -E\left\{\frac{\partial^2\ell_n(Y,\beta)}{\partial\beta^2}\right\} = X^{\mathrm{T}}\Sigma(X\beta)X, \tag{8}$$

and we write $A_n = A_n(\beta_{n,0})$. Thus A_n and B_n are the covariance matrices of $X^{\mathrm{T}}Y$ under the best working model $F_n(\beta_{n,0})$ and the true model G_n, respectively. To account for the effect of model misspecification, we define the covariance contrast matrix $H_n = A_n^{-1}B_n$ as in Lv & Liu (2014). Observe that A_n and B_n coincide when the best working model and the true model are the same. In this case, H_n is the identity matrix of size d.

2.2. *High-dimensional generalized Bayesian information criterion with prior probability*

Given a set of competing models $\{\mathfrak{M}_m : m = 1,\ldots,M\}$, a popular model selection procedure using the Bayesian principle of model selection involves first placing nonzero prior probability $\alpha_{\mathfrak{M}_m}$ on each model $\mathfrak{M}_m$, and then choosing a prior distribution $\mu_{\mathfrak{M}_m}$ for the parameter vector in the corresponding model. We use $d_m = |\mathfrak{M}_m|$ to denote the dimensionality of candidate model $\mathfrak{M}_m$ and suppress the subscript m for conciseness whenever confusion is unlikely. Assume that the density function of $\mu_{\mathfrak{M}_m}$ is bounded in $\mathbb{R}^{\mathfrak{M}_m} = \mathbb{R}^{d_m}$ and locally bounded away from zero in a shrinking neighbourhood of $\beta_{n,0}$. The Bayesian principle of model selection entails choosing the most probable model a posteriori, i.e., choosing the model $\mathfrak{M}_{m_0}$ such that

$m_0 = \arg\max_{m\in\{1,\ldots,M\}} S(y, \mathfrak{M}_m; F_n)$, where

$$S(y, \mathfrak{M}_m; F_n) = \log \int \alpha_{\mathfrak{M}_m} \exp\{\ell_n(y, \beta)\}\, \mathrm{d}\mu_{\mathfrak{M}_m}(\beta), \tag{9}$$

with the loglikelihood $\ell_n(y, \beta)$ defined in (4) and the integral over $\mathbb{R}^{d_m}$.

The choice of prior probabilities $\alpha_{\mathfrak{M}_m}$ is important in high dimensions. Lv & Liu (2014) suggested using prior probability $\alpha_{\mathfrak{M}_m} \propto \exp(-D_m)$ for each candidate model $\mathfrak{M}_m$, where the quantity D_m is defined as $D_m = E[I\{g_n; f_n(\cdot\,, \hat{\beta}_{n,m})\} - I\{g_n; f_n(\cdot\,, \beta_{n,m,0})\}]$, with the subscript m indicating a particular candidate model. The idea is that the further away the maximum likelihood estimator $\hat{\beta}_{n,m}$ is from the best misspecified generalized linear models $F_n(\cdot\,, \beta_{n,m,0})$, the lower the prior probability we assign to that model. In the high-dimensional setting where the dimensionality p can be much larger than the sample size n, it is sensible to also take into account the complexity of the space of all possible sparse models of the same size as $\mathfrak{M}_m$. This motivates us to consider a new prior probability of the form

$$\alpha_{\mathfrak{M}_m} \propto p^{-d} \exp(-D_m) \tag{10}$$

with $d = |\mathfrak{M}_m|$. The complexity factor p^{-d} is motivated by the asymptotic expansion of $\{p!/(p-d)!\}^{-1}$. In fact, an application of Stirling's formula yields $\log\{p!/(p-d)!\}^{-1} \approx -d\log p = \log(p^{-d})$ up to an additive term of order $o(d)$ when $d = o(p)$. The factor of $[p!/\{(p-d)!d!\}]^{-1}$ was also exploited by Chen & Chen (2008), who showed that by using the term $[p!/\{(p-d)!d!\}]^{-\gamma}$ with some constant $0 < \gamma \leqslant 1$, their extended Bayesian information criterion, EBIC, can be model selection consistent for the scenario of correctly specified models with $p = O(n^\kappa)$ for some positive constant κ satisfying $1 - (2\kappa)^{-1} < \gamma$. A different way of integrating the number of candidate models into the prior was considered by Szulc (2012) in the case where the model under study is correctly specified. Moreover, we add the term $d!$ to reflect a stronger prior on model sparsity. See also Fan & Tang (2013) for the characterization of model selection in ultrahigh dimensions with correctly specified models.

A similar normalization term can be found in some fully Bayesian methods; see, for instance, Castillo et al. (2015) for more details. However, the fully Bayesian methods need to specify the distribution of parameter β, whereas our method only puts some prior probabilities on the candidate models $\mathfrak{M}_m$, and the distribution $\mu_{\mathfrak{M}_m}(\beta)$ of parameter β given model $\mathfrak{M}_m$ does not need to be specified. Furthermore, fully Bayesian approaches require posterior computation, which may limit their use in high dimensions; see, for example, George (2000).

The asymptotic expansion in § 3.2 of the posterior model probability in Theorem 1 motivates us to introduce the high-dimensional generalized Bayesian information criterion with prior probability, $\mathrm{HGBIC_p}$, for large-scale model selection with misspecification.

DEFINITION 1. *Define* $\mathrm{HGBIC_p} = \mathrm{HGBIC_p}(y, \mathfrak{M}_m; F_n)$ *of model* $\mathfrak{M}_m$ *by*

$$\mathrm{HGBIC_p} = -2\ell_n(y, \hat{\beta}_n) + 2(\log p^*)|\mathfrak{M}_m| + \mathrm{tr}(\hat{H}_n) - \log|\hat{H}_n|, \tag{11}$$

where $\hat{H}_n$ *is a consistent estimator of* H_n *and* $p^* = pn^{1/2}$. *Here, consistency is in terms of trace and log determinant of the matrix.*

In correctly specified models, $H_n = A_n^{-1}B_n = I_d$ and so the term $\mathrm{tr}(\hat{H}_n) - \log|\hat{H}_n|$ in (11) is asymptotically close to $|\mathfrak{M}_m|$ when $\hat{H}_n$ is a consistent estimator of H_n. Thus, compared to the

BIC with factor $\log n$, the HGBIC_p contains a larger factor of order $\log p$ when the dimensionality p grows nonpolynomially with the sample size n. This leads to a heavier penalty on model complexity, similar to that in Fan & Tang (2013).

As shown in Lv & Liu (2014), the HGBIC_p defined in (11) can also be viewed as a sum of three terms: the goodness of fit, model complexity and model misspecification; see Lv & Liu (2014) for more details. Furthermore, HGBIC_p is also related to Takeuchi's information criterion, $\text{TIC} = -2\ell_n(y, \hat{\beta}_n) + 2\,\text{tr}(\hat{H}_n)$ (Takeuchi, 1976), which contains a similar model misspecification term $\text{tr}(\hat{H}_n)$, but lacks any model complexity term.

Our new information criterion HGBIC_p represents an important extension of the model selection criterion $\text{GBIC}_\text{p} = -2\ell_n(y, \hat{\beta}_n) + (\log n)|\mathfrak{M}_m| + \text{tr}(\hat{H}_n) - \log|\hat{H}_n|$ in Lv & Liu (2014), which was proposed for the scenario of model misspecification with fixed dimensionality, by explicitly taking into account the high dimensionality of the whole feature space. Moreover, in view of (11) and the definition of p^*, HGBIC_p has an additional model complexity term $2(\log p)|\mathfrak{M}_m|$.

3. Asymptotic properties of HGBIC_p

3.1. *Technical assumptions*

We list the technical assumptions required to prove the main results and the asymptotic properties of the maximum likelihood estimator with diverging dimensionality. Denote by Z the full design matrix of size $n \times p$ whose (i,j)th entry is x_{ij}. For any subset $\mathfrak{M}_m$ of $\{1, \ldots, p\}$, $Z_{\mathfrak{M}_m}$ denotes the submatrix of Z formed by the columns whose indices are in $\mathfrak{M}_m$. When confusion is unlikely, we drop the subscript and write $X = Z_{\mathfrak{M}_m}$ for fixed $\mathfrak{M}$. For theoretical reasons, we restrict the parameter space to $\mathcal{B}_0$, a sufficiently large convex and compact set of $\mathbb{R}^p$. We consider parameters with bounded support. Specifically, we define $\mathcal{B}(\mathfrak{M}_m) = \{\beta \in \mathcal{B}_0 : \text{supp}(\beta) = \mathfrak{M}_m\}$ and $\mathcal{B} = \bigcup_{|\mathfrak{M}_m| \leqslant K} \mathcal{B}(\mathfrak{M}_m)$, where the maximum support size K is taken to be $o(n)$. Moreover, we assume that $c_0 \leqslant b''(Z\beta) \leqslant c_0^{-1}$ for any $\beta \in \mathcal{B}$, where c_0 is some positive constant.

We use the following notation. For matrices, $\|\cdot\|_2$, $\|\cdot\|_\infty$ and $\|\cdot\|_\text{F}$ denote the matrix operator norm, entrywise maximum norm and matrix Frobenius norm, respectively. For vectors, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the vector L_2-norm and maximum norm, and $(v)_i$ represents the ith component of vector v. Denote by $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ the smallest and largest eigenvalues of a given matrix, respectively.

Assumption 1. There exists some positive constant c_1 such that for each $i = 1, \ldots, n$, $\text{pr}(|W_i| > t) \leqslant c_1 \exp(-c_1^{-1} t)$ for any $t > 0$, where $W = (W_1, \ldots, W_n)^\text{T} = Y - E(Y)$. The variances of the Y_i are bounded below uniformly in i and n.

Assumption 2. Let u_1 and u_2 be some positive constants and $\tilde{m}_n = O(n^{u_1})$ a diverging sequence. We have the bounds $\max\{\|E(Y)\|_\infty, \sup_{\beta \in \mathcal{B}} \|\mu(Z\beta)\|_\infty\} \leqslant \tilde{m}_n$ and $\sum_{i=1}^n \big([E(Y_i) - \{\mu(X\beta_{n,0})\}_i]^2 / \text{var}(Y_i)\big)^2 = O(n^{u_2})$. For simplicity, we also assume that $\tilde{m}_n$ diverges faster than $\log n$.

Assumption 3. Let $K = o(n)$ be a positive integer. There exist positive constants c_2 and u_3 such that for any $\mathfrak{M}_m \subset \{1, \ldots, p\}$ with $|\mathfrak{M}_m| \leqslant K$, we have $c_2 \leqslant \lambda_{\min}(n^{-1} Z_{\mathfrak{M}_m}^\text{T} Z_{\mathfrak{M}_m}) \leqslant \lambda_{\max}(n^{-1} Z_{\mathfrak{M}_m}^\text{T} Z_{\mathfrak{M}_m}) \leqslant c_2^{-1}$ and $\|Z\|_\infty = O(n^{u_3})$. For simplicity, we assume that the columns of Z are normalized, i.e., $\sum_{i=1}^n x_{ij}^2 = n$ for all $j = 1, \ldots, p$.

Assumption 1 is a standard tail assumption on the response variable Y; it ensures that the subexponential norm of the response is bounded. Assumptions 2 and 3 are counterparts of assumptions in Fan & Tang (2013). However, Assumption 2 has been modified to deal with model misspecification. More specifically, the means of the true distribution and fitted model, as well as their relations, are assumed in Assumption 2: the first part simultaneously controls the tail behaviour of the response and fitted model, while the second part ensures that the mean of the fitted distribution does not deviate from the true mean too significantly. We point out that such an assumption does not limit the generality of model misspecification, since the misspecification considered in this paper is due to the distributional mismatch between the working model and the underlying true model. Even in the misspecified scenario, the fitted mean vector from the working model can approximate the true mean vector under certain regularity conditions. Assumption 3 concerns the design matrix X; the first part is important for the consistency of the maximum likelihood estimator $\hat{\beta}_n$ and the uniqueness of the population parameter. Assumptions 2 and 3 also provide bounds on the eigenvalues of $A_n(\beta)$ and B_n. See Fan & Tang (2013) for further discussions of these assumptions.

For the next two assumptions, we define a neighbourhood around $\beta_{n,0}$. Let $\delta_n = \tilde{m}_n(\log p)^{1/2} = O\{n^{u_1}(\log p)^{1/2}\}$. We define the neighbourhood $N_n(\delta_n) = \{\beta \in \mathbb{R}^d : \|(n^{-1}B_n)^{1/2}(\beta - \beta_{n,0})\|_2 \leqslant (n/d)^{-1/2}\delta_n\}$. We assume that $(n/d)^{-1/2}\delta_n$ converges to zero so that $N_n(\delta_n)$ is an asymptotically shrinking neighbourhood of $\beta_{n,0}$.

Assumption 4. Assume that the prior density relative to the Lebesgue measure μ_0 on $\mathbb{R}^d$, $\pi\{h(\beta)\} = \mathrm{d}\mu_{\mathfrak{M}_m}/\mathrm{d}\mu_0\{h(\beta)\}$, satisfies $\inf_{\beta \in N_n(2\delta_n)} \pi\{h(\beta)\} \geqslant c_3$ and $\sup_{\beta \in \mathbb{R}^d} \pi\{h(\beta)\} \leqslant c_3^{-1}$, where c_3 is a positive constant and $h(\beta) = (n^{-1}B_n)^{1/2}\beta$.

Assumption 5. Let $V_n(\beta) = B_n^{-1/2}A_n(\beta)B_n^{-1/2}$, $V_n = V_n(\beta_{n,0}) = B_n^{-1/2}A_nB_n^{-1/2}$ and $\tilde{V}_n(\beta_1, \ldots, \beta_d) = B_n^{-1/2}\tilde{A}_n(\beta_1, \ldots, \beta_d)B_n^{-1/2}$, where $\tilde{A}_n(\beta_1, \ldots, \beta_d)$ is the matrix whose jth row is the corresponding row of $A_n(\beta_j)$ for each $j = 1, \ldots, d$. There exists some sequence $\rho_n(\delta_n)$ such that $\rho_n(\delta_n)\delta_n^2 d$ converges to zero, $\max_{\beta_1, \ldots, \beta_d \in N_n(\delta_n)} \|\tilde{V}_n(\beta_1, \ldots, \beta_d) - V_n\|_2 \leqslant \rho_n(\delta_n)$, and $\max_{\beta \in N_n(2\delta_n)} \max[|\lambda_{\min}\{V_n(\beta) - V_n\}|, |\lambda_{\max}\{V_n(\beta) - V_n\}|] \leqslant \rho_n(\delta_n)$.

Similar versions of Assumptions 4 and 5 were imposed in Lv & Liu (2014). Under Assumption 4, the prior density is bounded above globally and bounded below in a neighbourhood of $\beta_{n,0}$. This assumption is used in Theorem 1 for the asymptotic expansion of the posterior model probability. Assumption 5 is on the continuity of the matrix-valued function V_n and $\tilde{V}_n$ in a shrinking neighbourhood $N_n(2\delta_n)$ of $\beta_{n,0}$. The first and second parts control the expansions of the expected loglikelihood and score functions, respectively. Assumption 5 ensures that the remainders are negligible in approximating $S(y, \mathfrak{M}_m; F_n)$. More detailed discussion of Assumption 5 is provided in the Supplementary Material; see also Lv & Liu (2014) for further discussions about these assumptions.

3.2. *Asymptotic expansion of the Bayesian principle of model selection*

We now give the asymptotic expansion of the posterior model probability with the prior introduced in § 2.2. As mentioned earlier, the Bayesian principle chooses the model that maximizes $S(y, \mathfrak{M}_m; F_n)$ given in (9). To ease the presentation, for any $\beta \in \mathbb{R}^d$ we define a quantity

$$\ell_n^*(y, \beta) = \ell_n(y, \beta) - \ell_n(y, \hat{\beta}_n), \tag{12}$$

which is the deviation of the quasi-loglikelihood from its maximum. Then, from (9) and (12), we have

$$S(y, \mathfrak{M}_m; F_n) = \ell_n(y, \hat{\beta}_n) + \log E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n\} + \log \alpha_{\mathfrak{M}_m}, \tag{13}$$

where $U_n(\beta) = \exp\{n^{-1}\ell_n^*(y, \beta)\}$. With the choice of the prior probability in (10), it is clear that

$$\log \alpha_{\mathfrak{M}_m} = -D_m - d \log p. \tag{14}$$

Aided by (13) and (14), some delicate technical analysis unveils the following expansion of $S(y, \mathfrak{M}_m; F_n)$.

THEOREM 1. *Assume that Assumptions 1–5 hold and let* $\alpha_{\mathfrak{M}_m} = Cp^{-d}\exp(-D_m)$ *where* $C > 0$ *is a normalization constant. If* $(n/d)^{-1/2}\delta_n = o(1)$*, then with probability tending to* 1*,*

$$\begin{aligned} S(y, \mathfrak{M}_m; F_n) = \ell_n(y, \hat{\beta}_n) - (\log p^*)|\mathfrak{M}_m| - \frac{1}{2}\operatorname{tr}(H_n) + \frac{1}{2}\log|H_n| \\ + \log(Cc_4) + o(\tilde{\mu}_n), \end{aligned} \tag{15}$$

where $H_n = A_n^{-1}B_n$*,* $p^* = pn^{1/2}$*,* $\tilde{\mu}_n = \max\{\operatorname{tr}(A_n^{-1}B_n), 1\}$ *and* $c_3 \leqslant c_4 \leqslant c_3^{-1}$ *with* c_3 *being the positive constant in Assumption 4.*

Theorem 1 lays the foundation for investigating high-dimensional model selection with model misspecification. Based on the asymptotic expansion in (15), our new information criterion in (11) is defined by replacing the covariance contrast matrix H_n with a consistent estimator $\hat{H}_n$. The HGBIC_p naturally characterizes the impacts of both model misspecification and high dimensionality on model selection. A natural question is how to ensure a consistent estimator for H_n, which we address next.

3.3. *Consistency of covariance contrast matrix estimation*

For practical implementation of HGBIC_p, it is of vital importance to provide a consistent estimator for the covariance contrast matrix H_n. To this end, we consider the plug-in estimator $\hat{H}_n = \hat{A}_n^{-1}\hat{B}_n$ with $\hat{A}_n$ and $\hat{B}_n$ defined as follows. Since the maximum likelihood estimator $\hat{\beta}_n$ provides a consistent estimator of $\beta_{n,0}$ in the best misspecified generalized linear models $F_n(\cdot, \beta_{n,0})$, a natural estimate of the matrix A_n is

$$\hat{A}_n = A_n(\hat{\beta}_n) = X^{\mathrm{T}}\Sigma(X\hat{\beta}_n)X.$$

When the model is correctly specified, the simple estimator

$$\hat{B}_n = X^{\mathrm{T}}\operatorname{diag}\big[\{y - \mu(X\hat{\beta}_n)\} \circ \{y - \mu(X\hat{\beta}_n)\}\big]X,$$

with $\circ$ denoting the componentwise product, is an asymptotically unbiased estimator of the matrix B_n.

THEOREM 2. *Suppose that Assumptions 1–3 hold,* $n^{-1}A_n(\beta)$ *is Lipschitz in operator norm in the neighbourhood* $N_n(\delta_n)$*,* $d = O(n^{\kappa_1})$*, and* $\log p = O(n^{\kappa_2})$ *with constants satisfying* $0 < \kappa_1 < 1/4$*,* $0 < u_3 < 1/4 - \kappa_1$*,* $0 < u_2 < 1 - 4\kappa_1 - 4u_3$*,* $0 < u_1 < 1/2 - 2\kappa_1 - u_3$ *and* $0 < \kappa_2 < 1 - 4\kappa_1 - 2u_1 - 2u_3$*. Then the plug-in estimator* $\hat{H}_n = \hat{A}_n^{-1}\hat{B}_n$ *is such that* $\operatorname{tr}(\hat{H}_n) = \operatorname{tr}(H_n) + o_\text{P}(1)$

and $\log|\hat{H}_n| = \log|H_n| + o_{\mathrm{P}}(1)$ *with significant probability* $1 - O(n^{-\delta} + p^{1-8c_2\gamma_n^2})$, *where* δ *is some positive constant and* γ_n *is a slowly diverging sequence such that* $\gamma_n \tilde{m}_n (K^{1/2} n^{-1} \log p)^{1/2} \to 0$.

Theorem 2 improves upon the result of Lv & Liu (2014) in two important respects. First, the consistency of the covariance contrast matrix estimator was justified in Lv & Liu (2014) only in the scenario of correctly specified models. Our new result shows that the simple plug-in estimator $\hat{H}_n$ still enjoys consistency in the general setting of model misspecification. Second, the result in Theorem 2 holds for the case of high dimensionality. These theoretical guarantees are crucial for the practical implementation of our new information criterion. Our numerical studies in § 4 show that such an estimate works well in a variety of model misspecification settings.

3.4. *Model selection consistency of* HGBIC$_{\mathrm{p}}$

We further investigate the model selection consistency property of the information criterion HGBIC$_{\mathrm{p}}$. Assume that there are $M = o(n^{\delta})$ sparse candidate models $\mathfrak{M}_1, \ldots, \mathfrak{M}_M$, where δ is some sufficiently large positive constant. At first glance, such an assumption may seem slightly restrictive, since it rules out an exhaustive search over all $p!/\{(p-d)!d!\}$ possible candidate models. However, our goal here is to provide practitioners with some tools for comparing a set of candidate models that are available to them. In fact, the set of sparse models under model comparison can be often smaller in practice, for example polynomial instead of exponential in sample size, even in the ultrahigh-dimensional setting. One situation is where practitioners may apply different algorithms, each of which can lead to a possibly different model. Another example is the use of a certain regularization method with a sequence of sparse models generated by a path algorithm, which will be demonstrated in our numerical studies. For each candidate model $\mathfrak{M}_m$, we have the HGBIC$_{\mathrm{p}}$ criterion as defined in (11),

$$\mathrm{HGBIC}_{\mathrm{p}}(\mathfrak{M}_m) = -2\ell_n(y, \hat{\beta}_{n,m}) + 2(\log p^*)|\mathfrak{M}_m| + \mathrm{tr}(\hat{H}_{n,m}) - \log|\hat{H}_{n,m}|, \tag{16}$$

where $\hat{H}_{n,m}$ is a consistent estimator of $H_{n,m}$ and $p^* = pn^{1/2}$. Assume that there exists an oracle working model in the sequence $\{\mathfrak{M}_m : m = 1, \ldots, M\}$ that has support identical to the set of all important features in the true model. Without loss of generality, suppose that $\mathfrak{M}_1$ is such an oracle working model.

THEOREM 3. *Suppose that all the assumptions of Theorems 1 and 2 hold and that the population version of the* HGBIC$_{\mathrm{p}}$ *in* (16) *is minimized at* $\mathfrak{M}_1$ *such that for some positive sequence* Δ_n *slowly converging to zero,*

$$\min_{m>1} \left\{ \mathrm{HGBIC}^*_{\mathrm{p}}(\mathfrak{M}_m) - \mathrm{HGBIC}^*_{\mathrm{p}}(\mathfrak{M}_1) \right\} > \Delta_n \tag{17}$$

with $\mathrm{HGBIC}^*_{\mathrm{p}}(\mathfrak{M}_m) = -2\ell_n(y, \beta_{n,m,0}) + 2(\log p^*)|\mathfrak{M}_m| + \mathrm{tr}(H_{n,m}) - \log|H_{n,m}|$. *Then*

$$\min_{m>1} \left\{ \mathrm{HGBIC}_{\mathrm{p}}(\mathfrak{M}_m) - \mathrm{HGBIC}_{\mathrm{p}}(\mathfrak{M}_1) \right\} > \Delta_n/2$$

for large enough n with asymptotic probability 1.

Theorem 3 formally establishes the model selection consistency property of the new information criterion HGBIC$_{\mathrm{p}}$ for large-scale model selection with misspecification, in that the oracle working model can be selected from a large sequence of candidate sparse models with significant

probability. This desirable property is an important consequence of the results in Theorems 1 and 2. Furthermore, the assumption (17) is intrinsically necessary for this kind of theorem. For any model selection criterion, when the models are indistinguishable at the population level, the criterion cannot differentiate them in the sample version. Theorem 3 ensures that the gap in the population version is preserved in the sample version, giving a slight leeway.

4. Numerical studies

4.1. *Set-up*

In this section we investigate the finite-sample performance of the information criterion HGBIC_p in comparison with the information criteria AIC, BIC, EBIC (Chen & Chen, 2008), GIC (Fan & Tang, 2013), GAIC, or equivalently TIC, GBIC and GBIC_p in high-dimensional misspecified models via three simulation examples: a multiple-index model, a logistic regression model with interaction effects, and a Poisson regression model with interaction effects. For each candidate model $\mathfrak{M}_m$, the EBIC and GIC criteria are defined as

$$\text{EBIC}(\mathfrak{M}_m) = -2\ell_n(y, \hat{\beta}_{n,m}) + (\log n)|\mathfrak{M}_m| + \log\binom{p}{|\mathfrak{M}_m|},$$

$$\text{GIC}(\mathfrak{M}_m) = -2\ell_n(y, \hat{\beta}_{n,m}) + (\log n)(\log\log p)|\mathfrak{M}_m|.$$

4.2. *Multiple-index model*

The first model we consider is the multiple-index model

$$Y = f(\beta_1 X_1) + f(\beta_2 X_2 + \beta_3 X_3) + f(\beta_4 X_4 + \beta_5 X_5) + \varepsilon, \tag{18}$$

where the response depends on the covariates X_j only through the first five in a nonlinear fashion and $f(x) = x^3/(x^2+1)$. Here the rows of the $n \times p$ design matrix Z are sampled as independent copies from $N(0, I_p)$, and the n-dimensional error vector ε is distributed as $N(0, \sigma^2 I_n)$. We let the true parameter vector be $\beta_0 = (1, -1, 1, 1, -1, 0, \ldots, 0)^\text{T}$ and set $\sigma = 1$. We vary the dimensionality p from 100 to 3200 while keeping the sample size n fixed at 200. We aim to investigate the behaviour of different information criteria as the dimensionality increases. Although the data were generated from model (18), we fit the linear regression model (1). This is a typical example of model misspecification. Since the first five variables are independent of the other variables, the oracle working model is $M_0 = \text{supp}(\beta_0) = \{1, \ldots, 5\}$. Because of the high dimensionality, it is computationally prohibitive to implement the best subset selection. Therefore we first applied the lasso followed by least-squares refitting to build a sequence of sparse models and then selected the final model using a model selection criterion. In practice, one can apply any preferred variable selection procedure to obtain a sequence of candidate interpretable models.

We report the consistent selection probability, i.e., the proportion of simulations where the selected model $\hat{M}$ is equal to M_0, the sure screening probability (Fan & Fan, 2008; Fan & Lv, 2008), i.e., the proportion of simulations where $\hat{M} \supset M_0$, and the prediction error $E(Y - z^\text{T}\hat{\beta})^2$ where $\hat{\beta}$ is an estimate and (z, Y) is an independent observation for $z = (X_1, \ldots, X_p)^\text{T}$. To evaluate the prediction performance of different criteria, we calculated the average prediction error on an independent test sample of size 10 000. The results for the prediction error and model selection performance are summarized in Table 1. We also calculate the average number of false positives for each method in Table 2.

From Table 1 we observe that as the dimensionality p increases, the consistent selection probability tends to decrease for all criteria except the newly proposed HGBIC_p, which maintains at

Table 1. *Average results over* 100 *repetitions for the example in* § 4.2 *with all entries multiplied by* 100

p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC_p	HGBIC_p	Oracle
Consistent selection probability with sure screening probability in parentheses									
100	0 (100)	29 (100)	70 (100)	66 (100)	0 (100)	33 (100)	57 (100)	100 (100)	100 (100)
200	0 (100)	6 (100)	57 (100)	59 (100)	0 (100)	9 (100)	32 (100)	99 (100)	100 (100)
400	0 (100)	1 (100)	57 (100)	68 (100)	0 (100)	3 (100)	13 (100)	99 (100)	100 (100)
800	0 (100)	0 (100)	51 (100)	64 (100)	0 (100)	0 (100)	10 (100)	98 (100)	100 (100)
1600	0 (100)	0 (100)	39 (100)	59 (100)	0 (100)	0 (100)	9 (100)	98 (100)	100 (100)
3200	0 (100)	0 (100)	43 (100)	64 (100)	0 (100)	0 (100)	4 (100)	95 (99)	100 (100)
Mean prediction error with standard error in parentheses									
100	151 (2)	126 (2)	122 (1)	122 (1)	137 (2)	126 (2)	123 (1)	119 (1)	119 (1)
200	166 (2)	131 (2)	121 (1)	121 (1)	139 (2)	130 (2)	124 (1)	117 (1)	117 (1)
400	181 (3)	140 (2)	124 (1)	123 (1)	146 (2)	139 (2)	129 (2)	120 (1)	119 (1)
800	187 (2)	149 (2)	127 (1)	125 (1)	151 (2)	147 (2)	136 (2)	121 (1)	121 (1)
1600	185 (2)	154 (2)	128 (2)	124 (1)	152 (2)	152 (2)	137 (2)	119 (1)	119 (1)
3200	178 (2)	151 (2)	123 (1)	120 (1)	146 (2)	150 (2)	134 (2)	117 (1)	116 (1)

AIC, Akaike information criterion; BIC, Bayesian information criterion; EBIC, extended Bayesian information criterion of Chen & Chen (2008); GIC, generalized information criterion; GAIC, generalized Akaike information criterion; GBIC, generalized Bayesian information criterion; GBIC_p, generalized Bayesian information criterion with prior probability; HGBIC_p, high dimensional generalized Bayesian information criterion with prior probability.

Table 2. *Average false positives over* 100 *repetitions for the example in* § 4.2

p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC_p	HGBIC_p
100	15.35	1.84	0.49	0.58	7.05	1.75	0.86	0.00
200	24.30	3.53	0.76	0.70	7.43	3.07	1.39	0.01
400	31.46	5.58	0.73	0.53	8.32	5.11	1.98	0.01
800	34.12	7.21	0.87	0.60	8.26	6.20	2.58	0.02
1600	34.41	8.74	1.23	0.56	7.65	7.58	3.12	0.02
3200	33.41	8.64	0.93	0.48	7.25	8.28	3.26	0.04

least 95% consistent selection probability for all dimensionalities considered. Generally speaking, GAIC was an improvement over AIC, and GBIC and GBIC_p performed better than BIC in terms of both prediction and variable selection. The high-dimensional information criteria EBIC and GIC outperformed the traditional AIC and BIC. In particular, the model selected by our new information criterion HGBIC_p delivered the best performance with the smallest prediction error and highest consistent selection probability across all settings.

An interesting observation comes from comparing GBIC_p, GIC and HGBIC_p in terms of the model selection consistency property. While GBIC_p is comparable to HGBIC_p when the dimensionality is not large, such as $p = 100$, the difference between these two methods increases as the dimensionality increases. In the case where $p = 3200$, HGBIC_p has 95% success for consistent selection, while GBIC_p has a success rate of only 4%. This confirms the necessity of including the $\log p^*$ factor with $p^* = pn^{1/2}$ in the model selection criterion to take into account the high dimensionality, which is in line with the results of Fan & Tang (2013) for the case of correctly specified models. On the other hand, due to the lack of consideration of model misspecification, GIC is still outperformed by the newly proposed HGBIC_p across all dimensionalities considered.

We further study a family of model selection criteria induced by the HGBIC_p and characterized as follows:

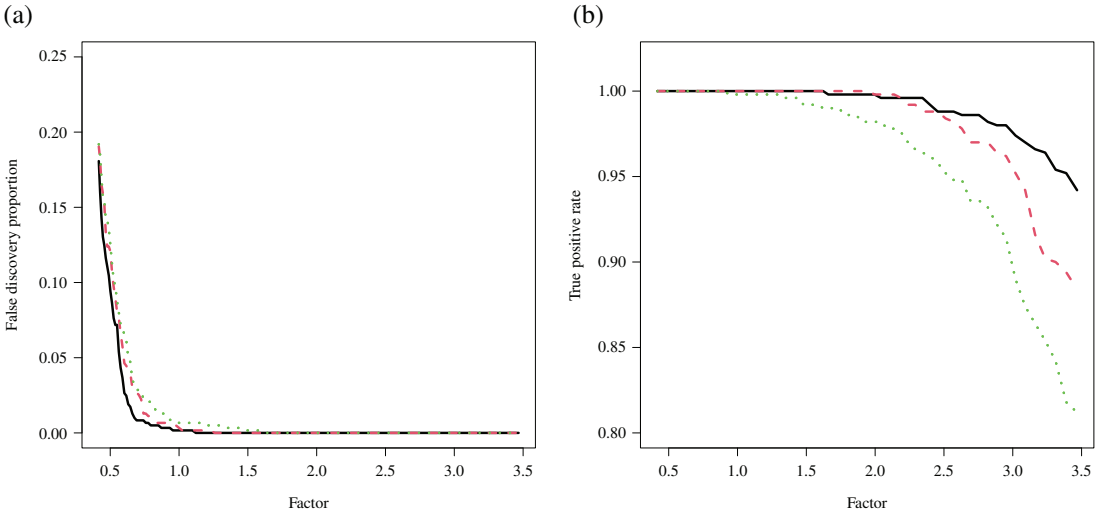

Fig. 1. For the example in § 4.2, (a) the average false discovery proportion and (b) the true positive rate as the factor ζ varies, when $p = 200$ (black solid), $p = 800$ (red dashed) and $p = 3200$ (green dot-dash).

$$\text{HGBIC}_{\text{p},\zeta}(\mathfrak{M}_m) = -2\ell_n(y, \hat{\beta}_{n,m}) + \zeta\{2(\log p^*)|\mathfrak{M}_m| + \text{tr}(\hat{H}_{n,m}) - \log|\hat{H}_{n,m}|\},$$

where ζ is a positive factor controlling the penalty level on both model misspecification and high dimensionality. The $\text{HGBIC}_{\text{p},\zeta}$ with $\zeta = 1$ reduces to our original HGBIC_{p}. Here we examine the impact of the factor ζ on the false discovery proportion and the true positive rate for the selected model $\hat{M}$, compared to the oracle working model M_0. In Fig. 1 we see that as ζ increases, the average false discovery proportion drops sharply as it gets close to 1. In addition, we have the desired model selection consistency property, with the false discovery proportion close to 0 and true positive rate close to 1 when $\zeta \in [1, 1.5]$. This figure demonstrates the robustness of the introduced $\text{HGBIC}_{\text{p},\zeta}$ criteria.

Acknowledgement

The authors sincerely thank the editor, associate editor and referees for comments that have helped to improve the paper significantly. This work was supported by the U.S. National Science Foundation, the Simons Foundation, and an Adobe Data Science Research Award. Demirkaya and Feng contributed equally to this work.

Supplementary material

Supplementary Material available at *Biometrika* online contains additional numerical studies, examples to compute HGBIC_{p}, all the proofs of the main results, and additional technical details.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B. N. Petrov & F. Csaki, eds. Budapest: Akademiai Kiado, pp. 267–81.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Auto. Contr.* **19**, 716–23.

Barber, R. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–85.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345–70.

BÜHLMANN, P. & VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin: Springer.
BÜHLMANN, P. & VAN DE GEER, S. (2015). High-dimensional inference in misspecified linear models. *Electron. J. Statist.* **9**, 1449–73.
CANDÈS, E. J., FAN, Y., JANSON, L. & LV, J. (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc.* B **80**, 551–77.
CASTILLO, I., SCHMIDT-HIEBER, J. & VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43**, 1986–2018.
CHEN, J. & CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–71.
CHEN, K. & CHAN, K.-S. (2011). Subset ARMA selection via the adaptive lasso. *Statist. Interface* **4**, 197–205.
CULE, M., SAMWORTH, R. & STEWART, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density (with Discussion). *J. R. Statist. Soc.* B **72**, 545–607.
EGUCHI, S. (2017). Model comparison for generalized linear models with dependent observations. *Economet. Statist.* **5**, 171–88.
FAN, J. & FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605–37.
FAN, J. & LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with Discussion). *J. R. Statist. Soc.* B **70**, 849–911.
FAN, J. & LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101–48. Invited review article.
FAN, J. & LV, J. (2018). Sure independence screening. *Wiley StatsRef*, doi: 10.1002/9781118445112.stat08043. Invited review article.
FAN, Y., DEMIRKAYA, E., LI, G. & LV, J. (2020). RANK: Large-scale inference with graphical nonlinear knockoffs. *J. Am. Statist. Assoc.* **115**, 362–79.
FAN, Y., DEMIRKAYA, E. & LV, J. (2019). Nonuniformity of p-values can occur early in diverging dimensions. *J. Mach. Learn. Res.* **20**, 1–33.
FAN, Y. & TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc.* B **75**, 531–52.
FONG, E. & HOLMES, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika* **107**, 489–96.
FOSTER, D. & GEORGE, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–75.
GEORGE, E. I. (2000). The variable selection problem. *J. Am. Statist. Assoc.* **95**, 1304–8.
HALL, P. (1990). Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Prob. Theory Rel. Fields* **85**, 449–67.
HSU, H.-L., ING, C.-K. & TONG, H. (2019). On model selection from a finite family of possibly misspecified time series models. *Ann. Statist.* **47**, 1061–87.
ING, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Ann. Statist.* **35**, 1238–77.
KONISHI, S. & KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–90.
KULLBACK, S. & LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.
LAURITZEN, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.
LIU, W. & YANG, Y. (2011). Parametric or nonparametric? A parametricness index for model selection. *Ann. Statist.* **39**, 2074–102.
LV, J. & LIU, J. S. (2014). Model selection principles in misspecified models. *J. R. Statist. Soc.* B **76**, 141–67.
MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.
NINOMIYA, Y. & KAWANO, S. (2016). AIC for the lasso in generalized linear models. *Electron. J. Statist.* **10**, 2537–60.
PEARL, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Amsterdam: Elsevier.
PENG, H., YAN, H. & ZHANG, W. (2013). The connection between cross-validation and Akaike information criterion in a semiparametric family. *J. Nonparam. Statist.* **25**, 475–85.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.
SHAH, R. D. & BÜHLMANN, P. (2018). Goodness-of-fit tests for high dimensional linear models. *J. R. Statist. Soc.* B **80**, 113–35.
STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc.* B **39**, 44–7.
SZULC, P. (2012). Weak consistency of modified versions of Bayesian information criterion in a sparse linear regression. *Prob. Math. Statist.* **32**, 47–55.
TAKEUCHI, K. (1976). The distribution of information statistics and the criterion of goodness of fit of models. *Math. Sci.* **153**, 12–18.
WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

[*Received on* 18 *March* 2020. *Editorial decision on* 15 *January* 2021]